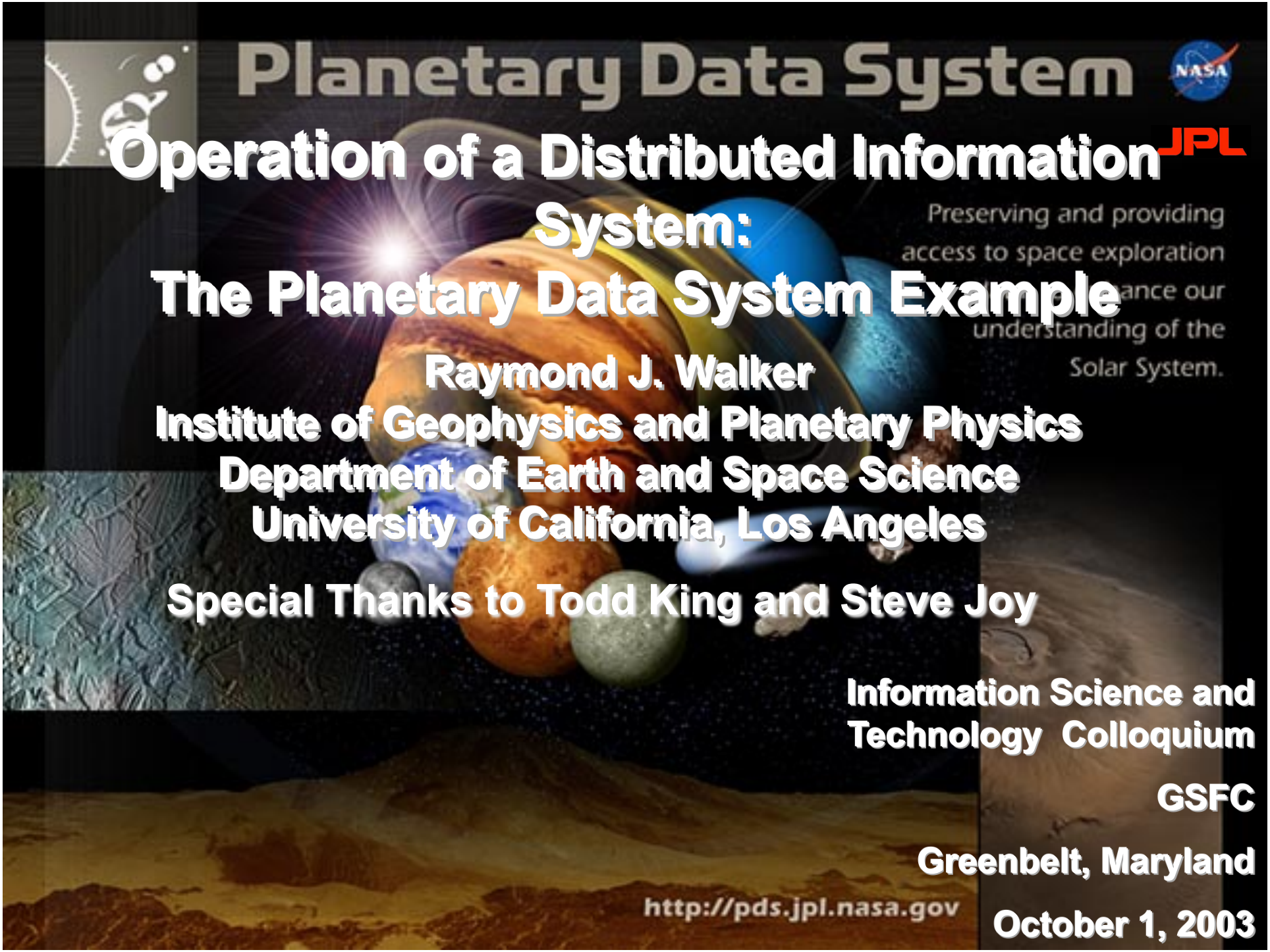


Operation of a Distributed Information System: The Planetary Data System Example

Raymond J. Walker
Institute of Geophysics and Planetary Physics
Department of Earth and Space Science
University of California, Los Angeles

Special Thanks to Todd King and Steve Joy

Information Science and Technology Colloquium

GSFC

Greenbelt, Maryland

October 1, 2003

The Beginning

- In the late 1970's the Committee on Data Management and Computation (CODMAC) of the National Research Council investigated the condition and accessibility of space derived data.
  - Data from NASA missions were being lost; were poorly documented and difficult to use; were difficult to access (*Bernstein et al.,* 1982).
  - Successful data management activities were run by scientists for scientists frequently without explicit government support (*Bernstein et al.* 1982).
- A second CODMAC study recommended a distributed approach to space science data management based on science considerations (*Arvidson et al*., 1986).

Solar System Exploration Gets Busy

- Under the leadership of Bill Quaide the planetary part of NASA responded to the NRC recommendations.
- A community wide workshop was held at Goddard Space Flight Center. It produced a two volume report that outlined what was to become PDS (*Keiffer et al.,* 1983a,b).
- The Pilot Planetary Data System was established at JPL. PPDS had heavy science involvement both through pilot distributed nodes and through the Planetary Science Data Steering Group (PSDSG).

The Development Phase

- Pilot Planetary Data System (Tom Renfrow, manager)
  - Developed the concept of data curation.
  - Developed documentation standards for the planetary data archive.
  - Developed a data catalog.
  - Carried out data restorations.
  - Nodes developed online data distribution systems. (*Note PDS has had both a distributed online archive and a distributable (CDROM/DVDROM) archive since its inception*.)
  - Severely criticized
    - Overly centralized.
    - Too little attention to data activities.

The PDS Goal

“Provide the scientific community with the highest quality planetary data forever!”

Sue McMahon

Why Bother?

- Many of the data are irreplaceable. They are national treasures.
- Analysis over long time intervals is frequently important.
- Having many scientists work with the data leads to innovation and maximizes science return.
- It is the law [e.g. *Dozier et al.,* 1995].

How PDS Works

- The structure of PDS
  - Why is PDS a distributed archive?
- The importance of standards
  - You can't create archival quality data without them.
- Working with data providers to create archival quality data
- Validating the data
  - The importance of peer review.
- Data Distribution

Why PDS is a Distributed Archive

- "*…it is clear that the most successful scientific use of space-acquired data occurs when interested scientists are actively involved in all elements of the data chain: planning, collecting, processing, archiving, distribution, analysis, and publication." Bernstein et al*., 1982 (NRC report).
- PDS is organized as a distributed data archive with active participation of scientists in the archiving processes.
  - PDS scientists work with the scientists who acquire the data to produce the archive.
  - PDS scientists help researchers access and use the archive.
- The best data products reside with people who use them to do their science.
- No single university or laboratory has the expertise necessary to understand all of data from all of the instruments involved in planetary science.

Discipline Nodes

- Defined by the scientific disciplines that make up planetary science.
  - geology and geophysics (Geo)
  - atmospheres (Atm)
  - plasma physics (PPI)
  - rings (Rings)
  - comets and asteroids (SBN)
- Provide expertise in each area of the science.
  - Assisted by subnodes that provide the required breadth.
  - Use temporary data nodes to assist with the archive of a given instrument or data set.
  - Provide continuity across missions.

Support Nodes

- Expert in a single field that transcends missions.
- Expertise required across disciplines.
  - Radio science (RS)
  - Imaging (Imaging)
  - Observation geometry and events (NAIF)
- Work with discipline nodes in the preparation and distribution of data sets.

Roles and Responsibilities

- Archive data sets from planetary missions.
  - Work with data providers to design data products.
  - Assist with the preparation of documentation to PDS standards.
  - Validate content and completeness of data sets and documentation.
- Distribute data to the science community.
  - Maintain detailed level catalogs (inventories).
- Work with science data users.
  - Provide expertise in using the data.
  - Help users produce special data products.
- Support education and public outreach activities.

Central Node

- Provides project management.
- Provides PDS interface to missions.
- Validates data products for standards compliance.
- Participates in defining standards.
- Maintains and documents PDS standards.
  - Planetary science data dictionary.
  - PDS standards document.
  - Data preparers workbook.
- Maintains high level planetary science data catalog.
- Supports education and outreach activities.

Examples of Science Node Involvement with Planetary Missions

	MGS	NEAR	Galileo	Cassini
Lead Node	Geo	SBN	PPI	Atm
Science Nodes	Atm, PPI	PPI	SBN, Rings, Atm	Rings, PPI, SBN
Support Nodes	Imaging, NAIF, RS			

Standards

Why Bother with Standards?

- Data have been lost for a number of reasons:
  - Inadequate documentation,
  - Poor data or documentation quality,
  - Outdated or unsupported formats,
  - Faulty or unsupported storage media.
- Standards help to insure that data are preserved in a usable form.
- Standards facilitate the use of data by scientists who were not involved with the data collection.

PDS Standards

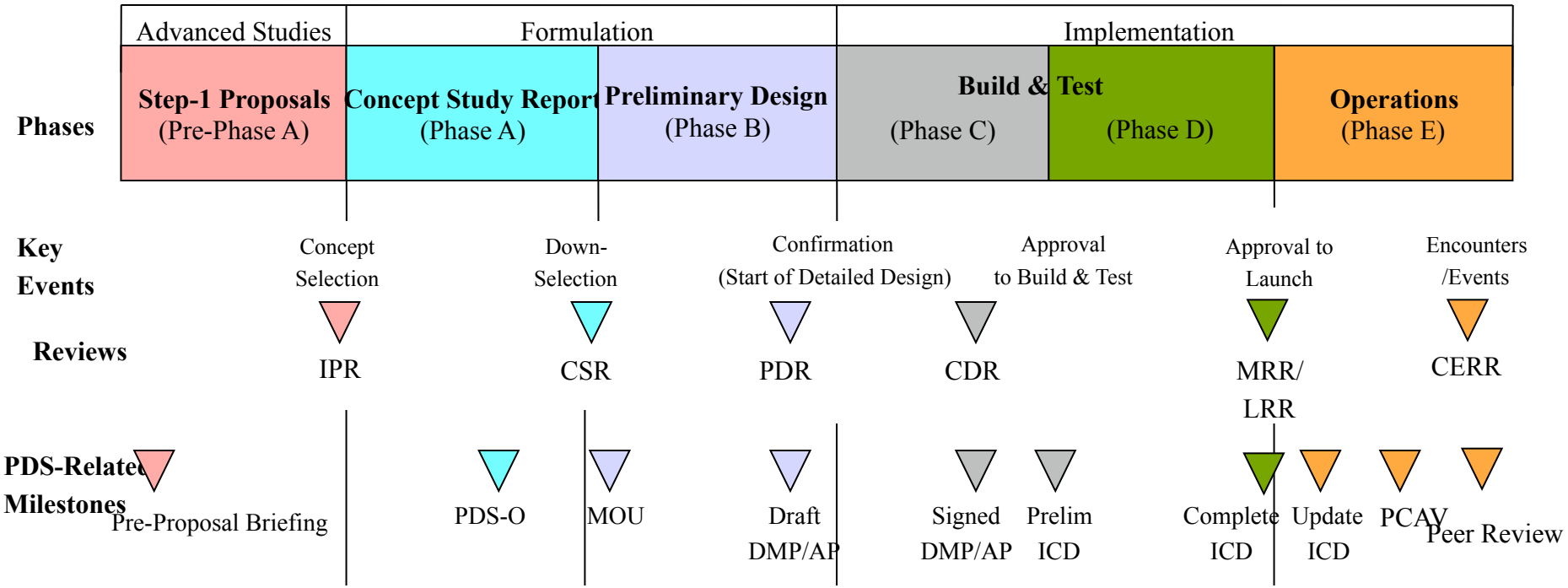
- Documentation.
  - Archive planning documents.
  - Description of missions and spacecraft.
  - Description of an instrument.
  - Description of the contents of a data set including geometry and calibration files.
  - Description of the format of a data file.
- Nomenclature.
  - PDS uses a *keyword* =*value* system to describe the data.
  - For this to work a dictionary of keyword definitions must be available to describe data and instruments.
- Organization.
  - Archive media.
  - Layout of archive volumes.
  - Required files and documentation.

Standard Organization of Archive Volumes

- PDS provides standards for how archive volumes are to be organized.
  - Standard top level directory structure.

  - Required “readme” files in each top level directory.
  - Labels required for all files on archive volume.
- Archive volumes may contain many data sets.
  - The Voyager 1 particles and fields data sets are collected on a single correlative science volume.
- Data sets can span many archive volumes.
  - The Voyager image collection spans may volumes.
- Archive volumes include ancillary information.
  - Calibration files, geometry information (SPICE), documents and software facilitate the use of the data.

PDS/Project Life Cycle



- CDR – Project/System Critical Design Review
- CERR – Critical Events (Encounters) Readiness Review
- CSR – Concept Study Report/Review
- DMP/AP – Data Management Plan /Archive Plan
- ICD – Interface Control Document
- IPR – Initial Proposal Review
- LRR – Launch Readiness Review
- MRR – Mission Readiness Review
- PCAV – PDS-Compliant Archive Volume
- PDR – Project/System Preliminary Design Review
- PDS-O – PDS Orientation

Working with Data Providers

- Meet with mission and investigator teams early in a mission and often during the mission.
  - The first meeting should be before the Project Data Management Plan (PDMP) is written.
  - Starting early saves work and reduces the cost in the long run.
  - Frequent meetings help to establish a good working relationship between the missions and PDS.
  - Frequent meetings help to reduce misunderstandings and errors.

The Project Data Management Plan

- PDMP is the blueprint for a successful archive.
  - PDS and the mission should jointly prepare the PDMP.
    - Writing the PDMP educates project management about the effort required to create a useful and long lasting archive.
    - Writing the PDMP educates PDS regarding mission science objectives and resources.
  - The PDMP should state the level of data products to be produced and why these are adequate archival products.
  - The PDMP should set out the schedule for delivery of products.
  - The PDMP should clearly specify the roles and responsibilities of the project, instrument teams, and the PDS.

Standard Data Product Descriptions

- The Interface Control Document (Software Interface Specification (SIS)) describes in detail the data to be delivered to the archive.
  - The purpose of a SIS is to describe the data set contents and structure to the PDS, and the science community in general.
  - A SIS is typically co-written by the PDS Node responsible for the data archive and the instrument team.
  - A SIS is written before any data are returned.
  - Writing a SIS requires investigators to think deeply about their data products.
    - Details of a SIS may be modified after receipt of data if data analysis indicates that adjustments are necessary.

Validating the Data
What is a PDS Peer Review?

- Each PDS data set undergoes a review modeled after the peer review used by scientific journals.
  - A review panel is selected by the PDS lead node.
  - The review panel consists of scientists who are experts in the use of similar data sets, members of the instrument team, and PDS standards experts.
  - Outside science experts are asked to try to use the data and documentation provided.
  - PDS experts review archive for compliance with PDS standards.

How are Data Sets Reviewed?

- The peer review process differs for different archive products.
  - Correlative science products (few volumes, many data sets) are reviewed by large panels (>15 people) for many weeks culminating in a formal review meeting.
  - High volume, wide distribution data sets (many volumes, few data sets) undergo a “volume design process” review.
    - The panel (5-15 people) reviews archive generation process and a characteristic sample of the data set from a limited number of initial volumes.
    - Subsequent volumes are validated by automated processes and PDS spot checks.
  - High volume, low distribution data sets undergo this same review method but typically involve fewer reviewers (4-5).

Does the review process delay getting the data out to the science community?

No!

- Data are typically made available to the science community as soon as the data can be put online. This helps to insure the quality of the final archive.
- Data under review are clearly marked as “in review”.
- Review panels are commonly selected from the list of scientists who acquire online data under review.

What happens after the review?

- The review panel points out problems with the data and documentation and suggests improvements.
  - Deficiencies are recorded as liens against the data set.
  - If changes to the data are requested, the instrument team will review the request for feasibility, budget impact, etc. There may be some negotiation between the review panel and the data provider regarding requested changes.
- When all outstanding liens against a data set are resolved, it is added to the PDS archive.
  - Catalog data are loaded into the PDS database.
  - Warnings are removed from online systems.
  - Final data are made available to the science community.
  - Copies are sent to NASA deep archive sites.
- PDS maintains errata lists for data sets.

Improving the Historical Archives

- In addition to including data from new and current missions, PDS has recovered data sets (some new) from historical missions.
  - Pioneer 10 and 11
    - High time resolution data acquired from the Pioneer 10 and 11 (P10/11) helium vector magnetometer for the Jupiter and Saturn flybys.
    - Recalibrated data from the P10/11 Geiger Tube Telescope recovering the saturated interval near Jupiter.
    - Full spectrum data from the P10/11 Plasma Analyzer.
  - Voyager
    - Improved pointing information for Voyager Jupiter and Saturn encounters.
    - High time resolution data from the Voyager Low Energy Charged Particle Experiment at Jupiter and Saturn.
    - Raw radio science data
  - Viking
    - Orbiter observation geometry information placed in SPICE formats.
    - Lander labeled release data.
  - Mariner 10
    - Restored data from the magnetometer and plasma (electron) instruments.
    - We were unable to make a useful data set for the energetic particle instrument.
- The cost of restoring data is much greater than the cost of initially preparing the data to PDS standards.

Data Distribution

- PDS has three (3) main user groups:
  - Planetary and Space Scientists
  - Mission Planners
  - Education and Public Outreach
- Each user groups places different, and sometimes conflicting, requirements on the system.

User Requirements

Science

- high resolution data
- both raw and processed data products
- rapid access to the latest data
- lots of data
- detailed documentation
- transformation and/ or data analysis support

Mission Planner

- processed or derived data products
- access to historic data sets
- detailed documentation
- expert assistance with preparing or analyzing data
- projections and modeling

EPO

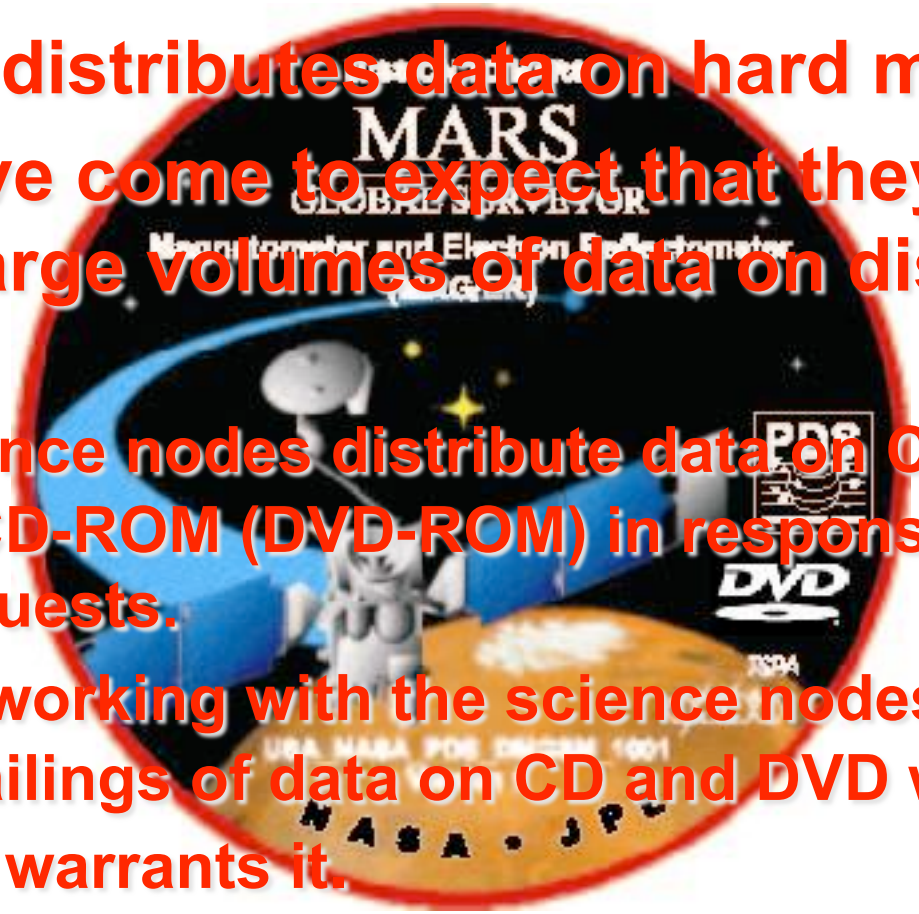
- small amounts of highly derived products (maps, plots, animations, etc.)
- current and historic data sets
- easy to understand documentation

Data Distribution

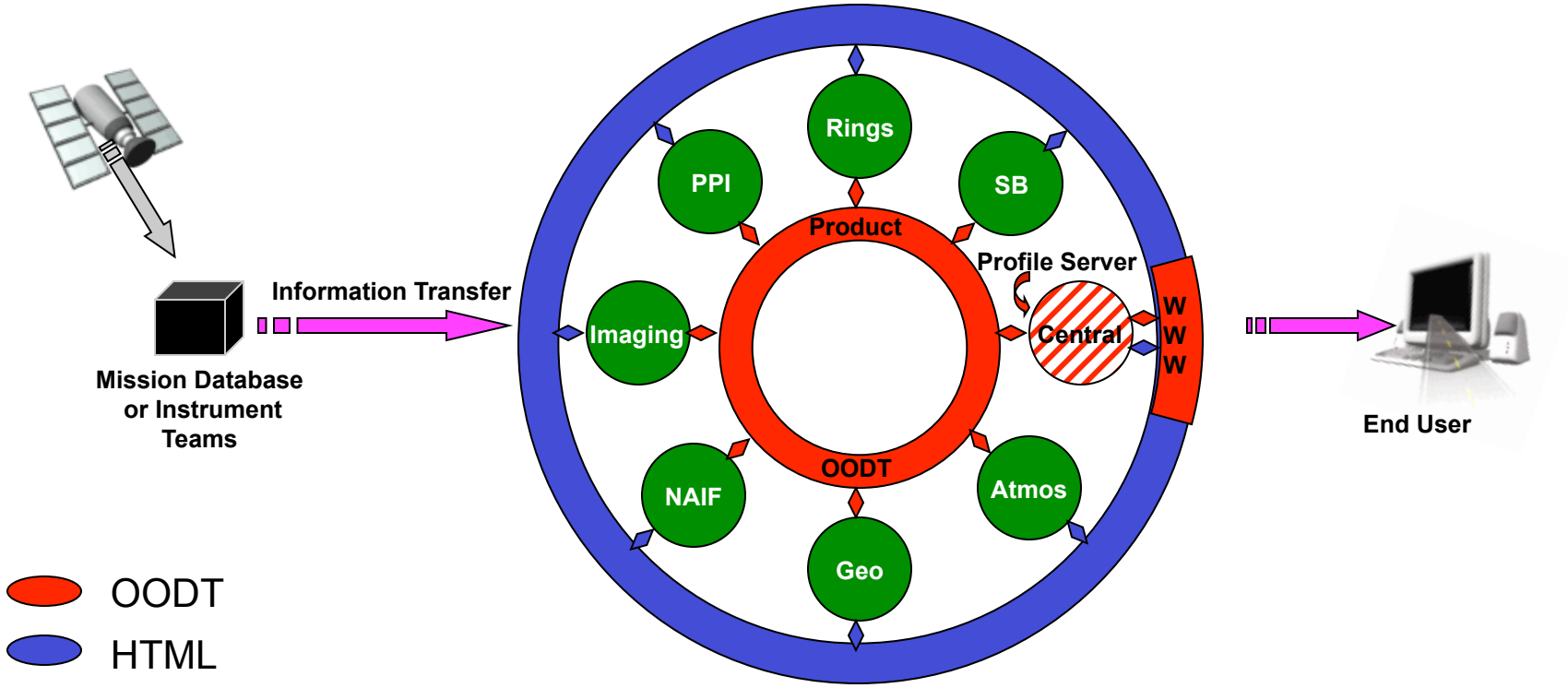
- PDS supports online data distribution.
  - Users can search the online catalog at the Central Node and order data sets.
    - Catalog data requests are directed to the Discipline Nodes where the orders are filled.
    - Originally only complete, archived data sets were available through catalog searches.
    - Starting with Mars Odyssey a new system called PDS-D allowed cross instrument searches and data extraction.
  - Users can search science node data inventory systems and place orders directly with the appropriate node.
    - Immediate online data delivery is supported by most nodes.
    - Data in preparation (under review) can be ordered from most nodes.
    - Subsets of data sets (individual files) can be ordered from most nodes.

Data Distribution Continued

- PDS also distributes data on hard media.
- Users have come to expect that they can request large volumes of data on distributable media.
  - The science nodes distribute data on CD-R (DVD-R) and CD-ROM (DVD-ROM) in response to specific user requests.
  - The CN working with the science nodes organizes large mailings of data on CD and DVD when demand warrants it.



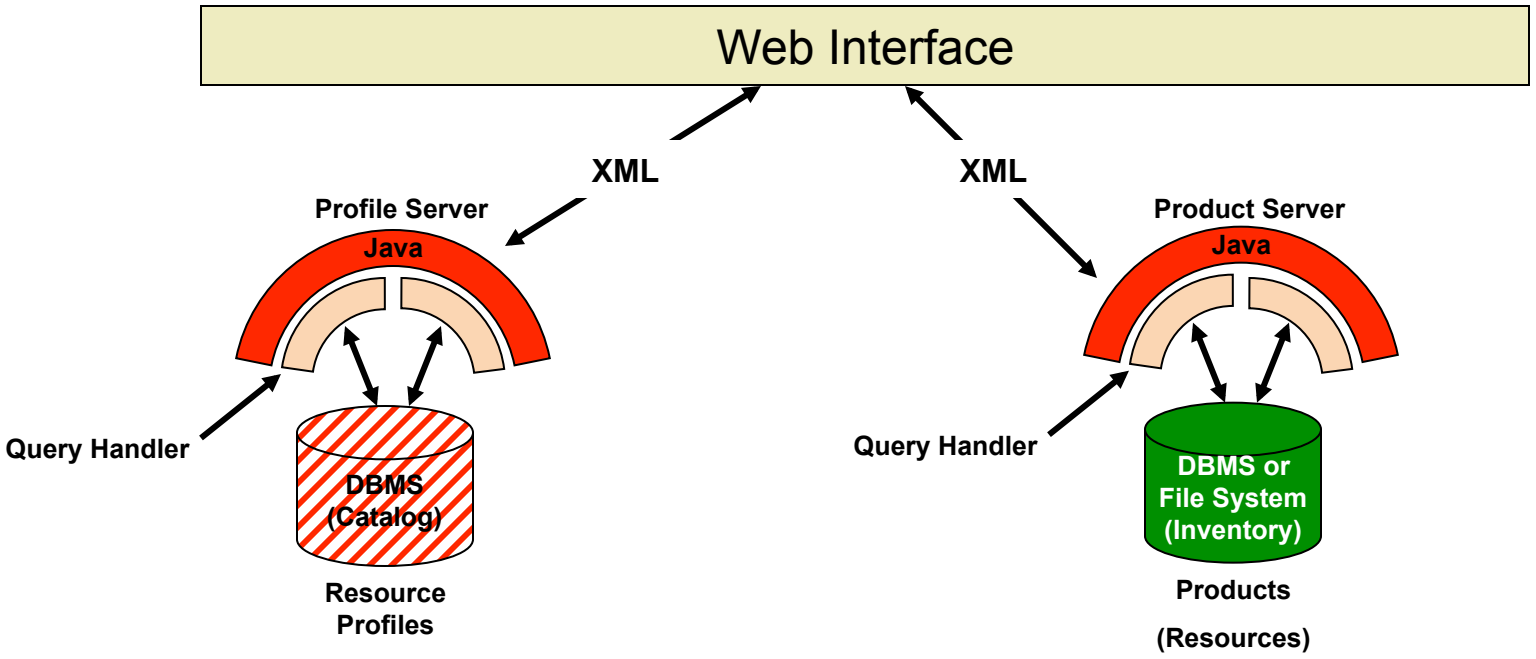
PDS Architecture



- OODT
- HTML
- Inventories, data access and data
- Catalog

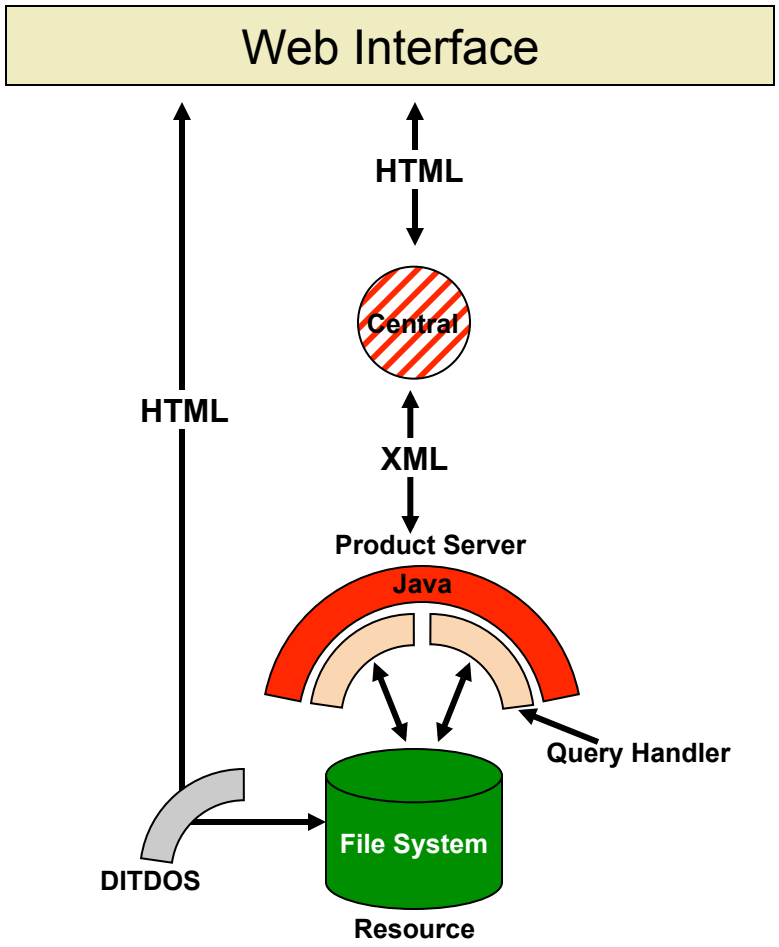
Object Oriented Data Technology

- Object Oriented Data Technology (OODT) is a web service.
- Queries and transfers are encapsulated in XML.
- Servers are analogous to a Java enabled web server that supports servlets (query handlers).



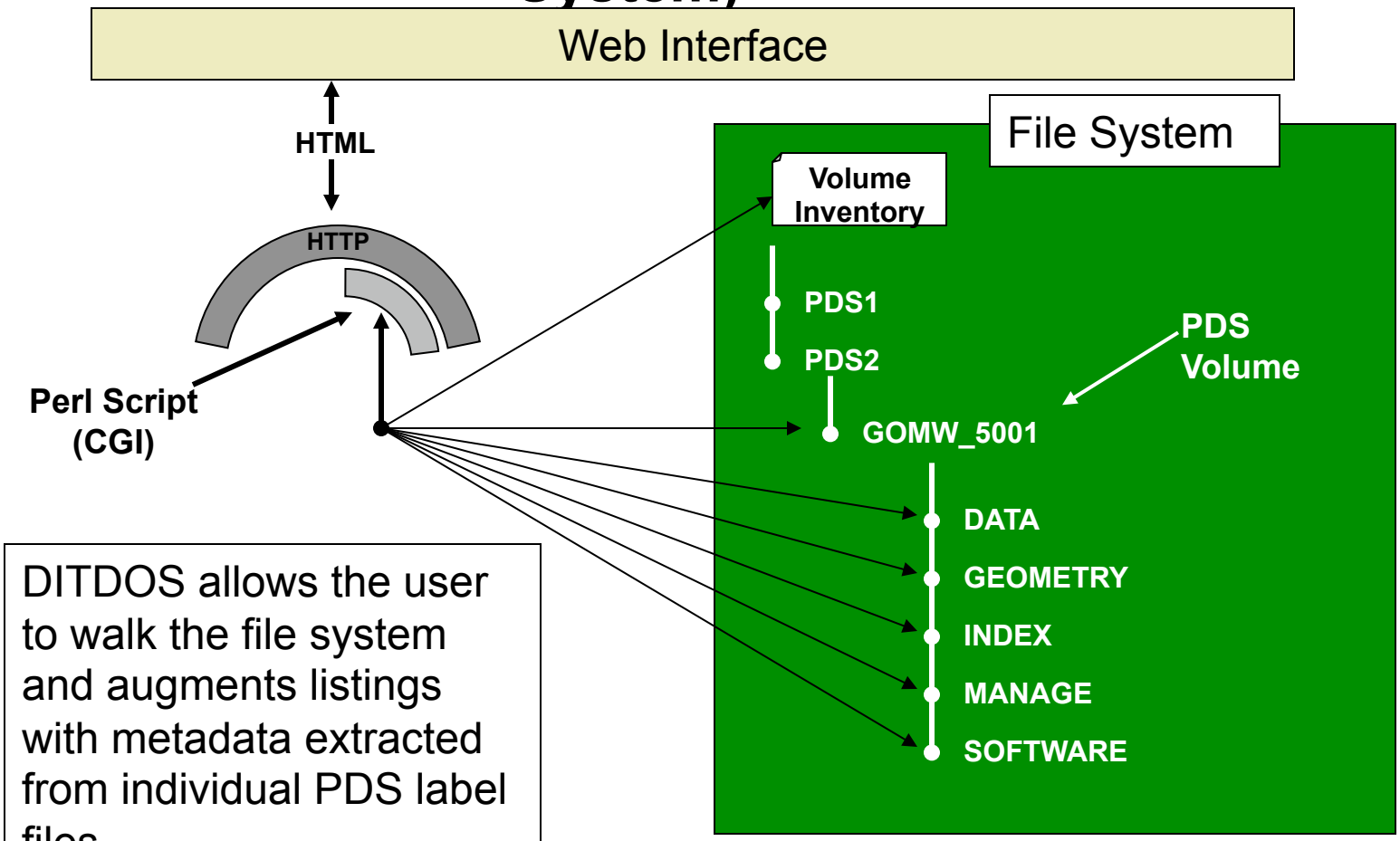
PDS/PPI Node - Today

- There are two ways to find and access the data.
  - –PDS-D
  - –A node developed system with an HTML interface to the Distributed Inventory Tracking and Data Ordering System.



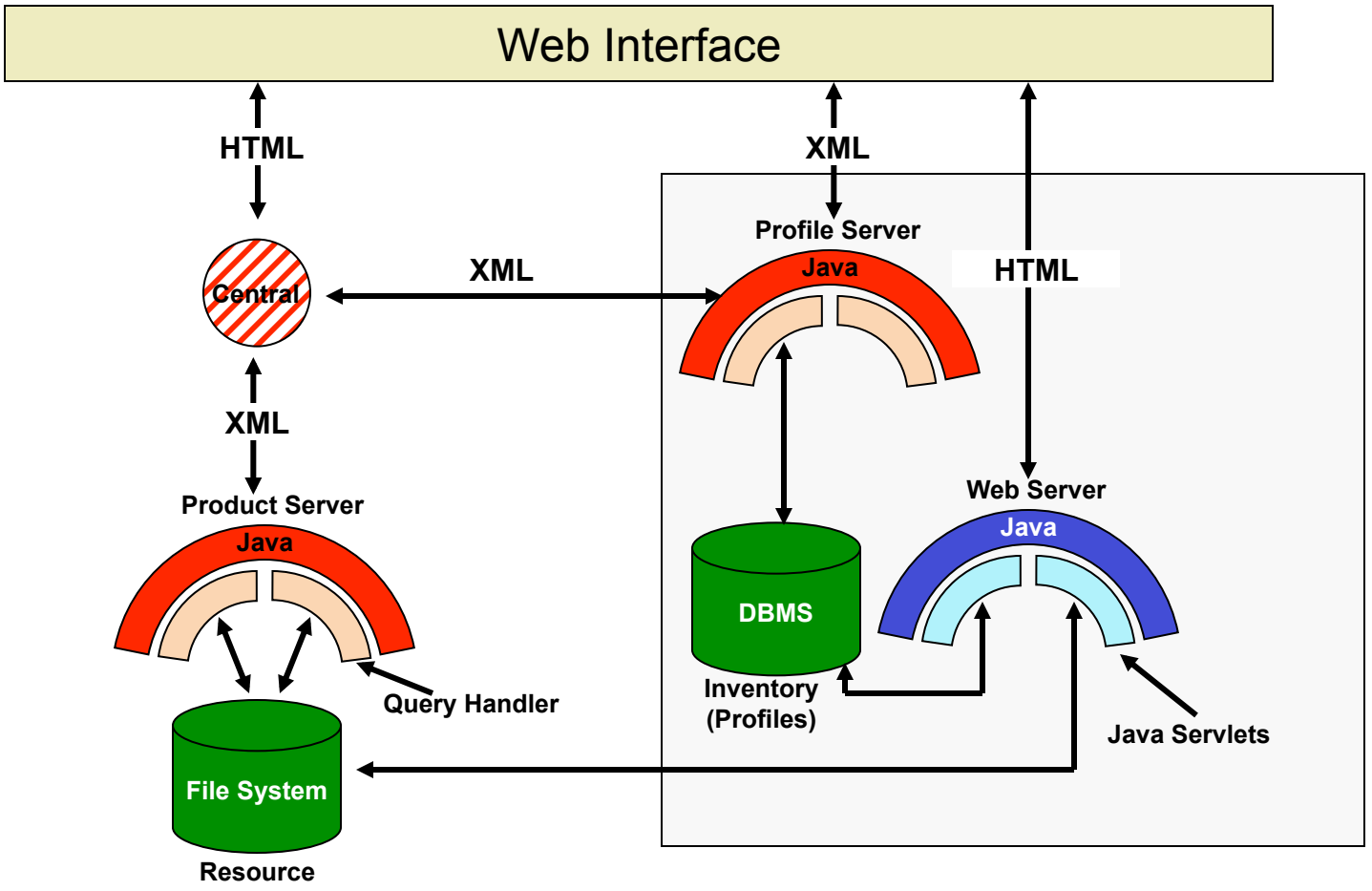
DITDOS

(Distributed Inventory Tracking and Data Ordering System)



DITDOS allows the user to walk the file system and augments listings with metadata extracted from individual PDS label files.

PDS/PPI Node- Future



Data Requests Requiring Special Handling

- Discipline Nodes have the science expertise necessary to fill special processing or handling data requests.
  - Requests to enhance the archive with value added products.
  - Requests to repackage data in the archive.
    - Provide data on alternative media.
    - Create special subsets or supersets of data.
    - Create ASCII files from binary data sets.
    - Merge data files.
  - Requests to transform or process archive data into new forms.
    - Resample data.
    - Transform data into new coordinate systems.
    - Apply calibrations or corrections to the data.

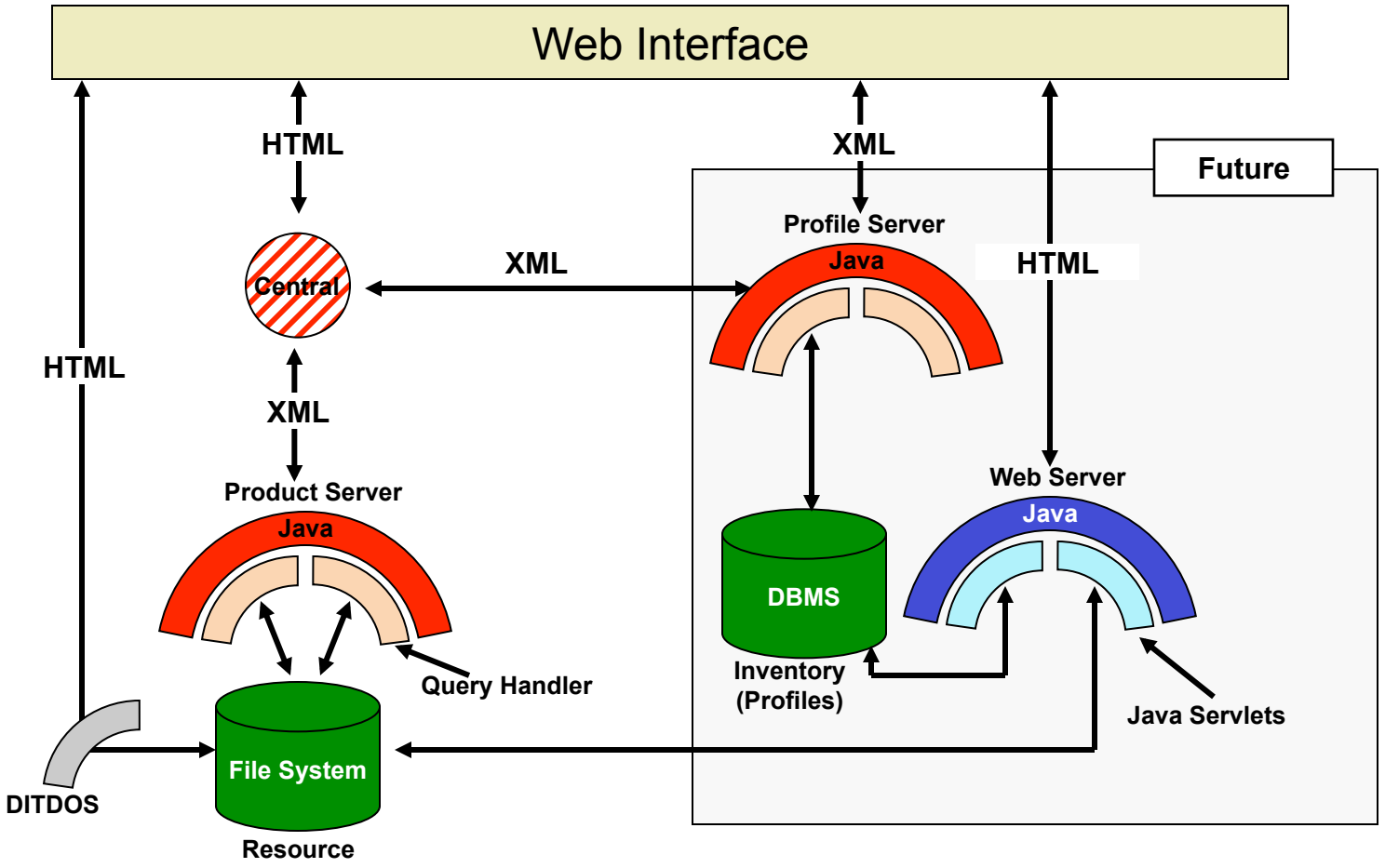
Future Challenges

- The next generation of missions will produce vast data volumes.
  - The Mars Reconnaissance Orbiter (MRO) will produce ~300TB of raw data during its nominal mission.
  - This is 100 times that of the current volume leader Mars Odyssey (~3TB).
- Distributing large data sets presents challenges to the PDS on top of those associated with acquiring, validating, and ingesting them.
  - Data rates are increasing faster than storage media capacity.
  - Hard media are costly to produce and distribute.
  - Electronic downloads may not be feasible for some users.
- We need better tools to select subsets of data for analysis from these large data volumes.

Future Challenges Continued

- The nature of the science is changing.
  - Discipline boundaries are becoming less distinct.
  - With increasing frequency investigations require data from more than one Node.
  - Researchers need to view the data with finer granularity than data sets when selecting data for an investigation.

PDS/PPI Node


Web Interface
HTML
XML
Future
Profile Server
Java
Central
XML
HTML
HTML
XML
Product Server
Java
Web Server
Java
DBMS
Query Handler
Inventory
(Profiles)
Java Servlets
DITDOS
File System
Resource